GOblet: Annotation of anonymous sequence data with Gene Ontology and Pathway terms

Detlef Groth[1,4], Stefanie Hartmann[1], Georgia Panopoulou[2], Albert J. Poustka[2] and Steffen Hennig[3]

[1]Potsdam University, Bioinformatics Group, c/o Max Planck Insitute of Molecular Plant Physiology,
Am Mühlenberg 1, D-14476 Potsdam-Golm, Germany

[2]Max Planck Institute for Molecular Genetics, Ihnestr. 63/73, D-14195 Berlin, Germany

[3]ImaGenes GmbH, Robert-Rössle-Str.10, D-13125 Berlin, Germany

Summary

The functional annotation of genomic data has become a major task for the ever-growing number of sequencing projects. In order to address this challenge, we recently developed GOblet, a free web service for the annotation of anonymous sequences with Gene Ontology (GO) terms. However, to overcome limitations of the GO terminology, and to aid in understanding not only single components but as well systemic interactions between the individual components, we have now extended the GOblet web service to integrate also pathway annotations. Furthermore, we extended and upgraded the data analysis pipeline with improved summaries, and added term enrichment and clustering algorithms. Finally, we are now making GOblet available as a stand-alone application for high-throughput processing on local machines. The advantages of this frequently requested feature is that a) the user can avoid restrictions of our web service for uploading and processing large amounts of data, and that b) confidential data can be analysed without insecure transfer to a public web server. The stand-alone version of the web service has been implemented using platform independent Tcl-scripts, which can be run with just a single runtime file utilizing the Starkit technology. The GOblet web service and the stand-alone application are freely available at `http://goblet.molgen.mpg.de`.

1 Introduction

The rapidly increasing number of sequencing projects poses a significant challenge for bioinformatics to extract relevant biological information from this wealth of high-throughput data. The experience with the recently finished human genome showed that the sequence information alone does not provide sufficient information to allow transformation of data into biology. When sequence data first becomes available, a common approach for annotation is a homology search for genes with known function and the transfer of their functional information. In a second step, genomic technologies like microarrays, proteomics, and metabolomics can then be used to improve knowledge about biological function by expanding and complementing the predictive annotation.

[4]To whom correspondence should be addressed. E-mail: `dgroth@mpimp-golm.mpg.de`

doi:10.2390/biecoll-jib-2008-104

It has recently been observed that often not individual genes, but larger structural units like biological networks and pathways are the driving forces of development and evolution [1]. Accordingly, we believe that for annotating gene function to new sequence data it is necessary to go beyond the single gene/protein view provided by pairwise sequence comparison tools like BLAST [2]. In this paper, we are introducing an approach for gene and genome annotation that integrates views provided by ontologies and metabolic or signalling pathways. Our approach thus takes into consideration system-level interactions of cellular components for the purpose of gene annotation.

1.1 Ontologies

Controlled vocabularies provide a way to make functional annotation accessible to computers. Furthermore, they are decisive for the integration of different datasets. The first controlled vocabulary was made by biochemists and has led to the Enzyme Commission Classification (EC, `http://www.chem.qmul.ac.uk/iubmb/enzyme/`), where enzymes are ordered in a four-level hierarchy. As an example, all oxidoreductases start with "1", more specific terms which Oxidoreductases "Acting on the CH-OH group of donors" have a "1.1" at the beginning, "1.1.1" means "with NAD+ or NADP+ as acceptor" and finally "1.1.1.1" are "alcohol dehydrogenase". In the last years a large number of biomedical ontologies where developed causing data integration problems due to different formats and structure of these ontologies. The Open Biomedical Ontologies (OBO) consortium with its OBO Foundry initiative tries to coordinate the development of ontologies. As a result there is an expanding family of ontologies providing high level of interoperability [3].

1.2 GO

One of the most widely used controlled vocabulary is the Gene Ontology (GO), which attempts to describe each gene product by its molecular function, the biological process in which it participates, and the cellular component in which it is localized. Currently, the GO contains about 14.000 biological process terms, about 8.000 molecular function terms, and about 2.000 cellular component terms [4], see Table 1.2. All terms are manually curated and connected by parent-child relationships expressed as "is_a" relations. The GO is organized as a directed acyclic graph, in which each term can be the child of more than one parent term. Annotation based on the structured vocabulary provided by GO-terms has become one of the standard methods to extract higher level biological meaning from experimental data [5, 6]. This is especially the case for gene expression studies, where data for a large number of genes needs to be explored routinely. An widely accepted tool, which integrates GO and microarray analysis is the Onto-Express suite of programs published by Draghici and coworkers [7], which was recently extended to account also for pathway information (`http://vortex.cs.wayne.edu/projects.htm`).

Despite its widespread use however, there are several limitations of the Gene Ontology and its usage. It contains, for instance, very different nesting levels for different branches of the GO hierarchy. This can lead to obscured statistical observations or to different quality and specificity of annotations in different branches (e.g., due to uneven contributions of the community) of the GO hierarchy. Another potential problem is that our current knowledge of

category	BP	MF	CC	BP	MF	CC
	2005			2008		
go-ids	8924	6929	1397	14659	8260	2064
obsoletes	330	526	113	471	566	117
MetaCyc	517	3460	0	680	3524	1
EC	1	4306	0	1	4762	0
slim-generic	51	41	36	53	42	36
slim-plant	52	28	29	51	27	27
slim-yeast	35	23	25	35	23	25

Table 1: GO-Statistics: Gene Ontology statistics for the three main categories, biological process (BP), molecular function (MF) and cellular component (CC). Data are taken from the obo files taken from the geneontology website for 2005 (January) and 2008 (April). Shown are the numbers of GO-ids, obsolete GO-ids, the number of GO-ids annotated with MetaCyc-terms or EC-numbers and the number of GO-ids for the GO-slims generic, plant and yeast.

biological features might not be completely expressed in terms of such a controlled vocabulary, in part because the vocabulary might be too restricted. The GO term for “extracellular region” (GO:0005576), for example, certainly is not a “cellular_component”, as the information in the GO graph currently suggests. Another problem is that the GO vocabulary was originally developed for the annotation of eukaryotic genomes. Therefore, the GO terminology might not be as accurate for prokaryotes as it is for eukaryotes. However, the recent addition of a prokaryotic subset to the ontology terminology has given an adjusted prokaryotic view on the ontology. For example in the prokaryotic subset terms like “mitochondrion” “nucleus” have now been excluded. Finally, because the GO hierarchy contains a very large number of GO terms, it is relatively difficult to summarize the results in an intuitive way. In order to reduce the number of GO terms, either a restriction of certain levels of the GO hierarchy (e.g., reporting only all children of “catalytic activity”) might be feasible, or so called GO slims (`http://www.geneontology.org/GO.slims.shtml`) can be used. The latter are cut-down versions of the GO ontologies, giving a broad overview of the ontology content without all the details of the specific fine grained terms. Slims have been used for example to provide an overview of the functional composition of proteomes [8]. An interesting alternative approach to Slim-based subsets on the GO hierarchy based on species data was recently introduced by Kusnierzcyk [9].

1.3 Pathways

Whereas the sequence of a given organism represents knowledge about its potential capabilities, the exploration of active pathways provides more information about the actual properties of the living cell. The simultaneous view on both sequence information and gene expression or metabolite data can therefore be a powerful tool to gain deeper insights into the status of an organism. Such combined and integrative approaches that go beyond the knowledge of single components are leading towards systems biology.

One of the oldest approaches to map pathway reactions onto structured vocabularies is the already mentioned EC Nomenclature. Important current pathway databases on the web are KEGG [10], Reactome [11], and MetaCyc [12]. The latter is a non-redundant, organism-independent database of small molecule metabolism and contains only experimentally verified data. MetaCyc-pathway and -reaction terms are structured like GO terms, i.e., they are organized in a hierarchical manner which can be explored online (`http://biocyc.org/META/class-tree?object=Pathways`). As the current GO database contains also mappings to EC identifiers and MetaCyc terms, the usage of GO annotations for pathway explorations is greatly simplified.

When sequences are annotated with pathway terms, the aforementioned limitations of the GO terminology usage might be overcome. One example for the annotation of sequences with KEGG terms is the stand-alone Python-based annotation tool provided by Mao and coworkers [13].

1.4 Annotation Summaries

After the successful annotation of sequence data with ontology and pathway terms, important questions should be addressed before starting computationally expensive and more detailed analyses. For example it is highly valuable to determine, the proportion of sequences that could be annotated, the most frequent annotation terms, and whether these are in fact more frequent than can be expected from their overall distribution. These so-called enrichment analyses of terms is generally done using reduced versions of the Gene Ontology tree, as many leaf nodes are too detailed and/or too rarely annotated to be useful to get an overview of the available annotations. A popular method for this kind of analysis is the Fisher Exact Test, originally proposed for 2x2 contingency tables [14]. The P-value of this test gives the probability with which the observed frequency in the data set could be generated by random extraction of genes from the reference data set. However, when testing multiple hypotheses in the same experiment the increased frequency of random events requires adjustment of the P-value. In the simplest case, each P-value is multiplied with the number of calculated P-values. This so-called Bonferroni correction is the most stringent correction and thereof used in our web service.

1.5 GOblet web server

The GOblet web server (`http://goblet.molgen.mpg.de`) is an automated BLAST-based annotation tool [15, 16]. It is free and open to all users without a login requirement. The web server accepts nucleotide and protein sequences in FASTA format and annotates them with Gene Ontology and pathway terms using BLAST-based sequence comparisons. Terms are tested for possible enrichments in comparison to the selected reference dataset. The Tcl scripting language (`http://www.tcl.tk/`) is used as the general programming language gluing together different parts of the application. C-coded functions are used for areas with critical processing speed, like database searches and BLAST-result parsing. The programming language and the libraries were chosen in order to simplify custom installations by downloading a stand-alone version. The GOblet has been running now for more than five years. The number of sequence submissions typically ranges between one and more than three thousands request per year, each submitting up to 500kb of nucleotide or protein sequences. In 2007, for example, almost 100.000 sequences were submitted by more than 120 different IP- addresses outside of the authors' institutions. In this paper we introduce a major upgrade of the GOblet web service. Specifically, we have now extended the GOblet web service to integrate also pathway annotations. Furthermore, we have extended and upgraded the data analysis pipeline with improved summaries, and we have added term enrichment and clustering algorithms. Finally, we are now making GOblet available as a stand-alone application for high-throughput processing on user's local machines.

2 Implementation and Usage

2.1 Data preparation

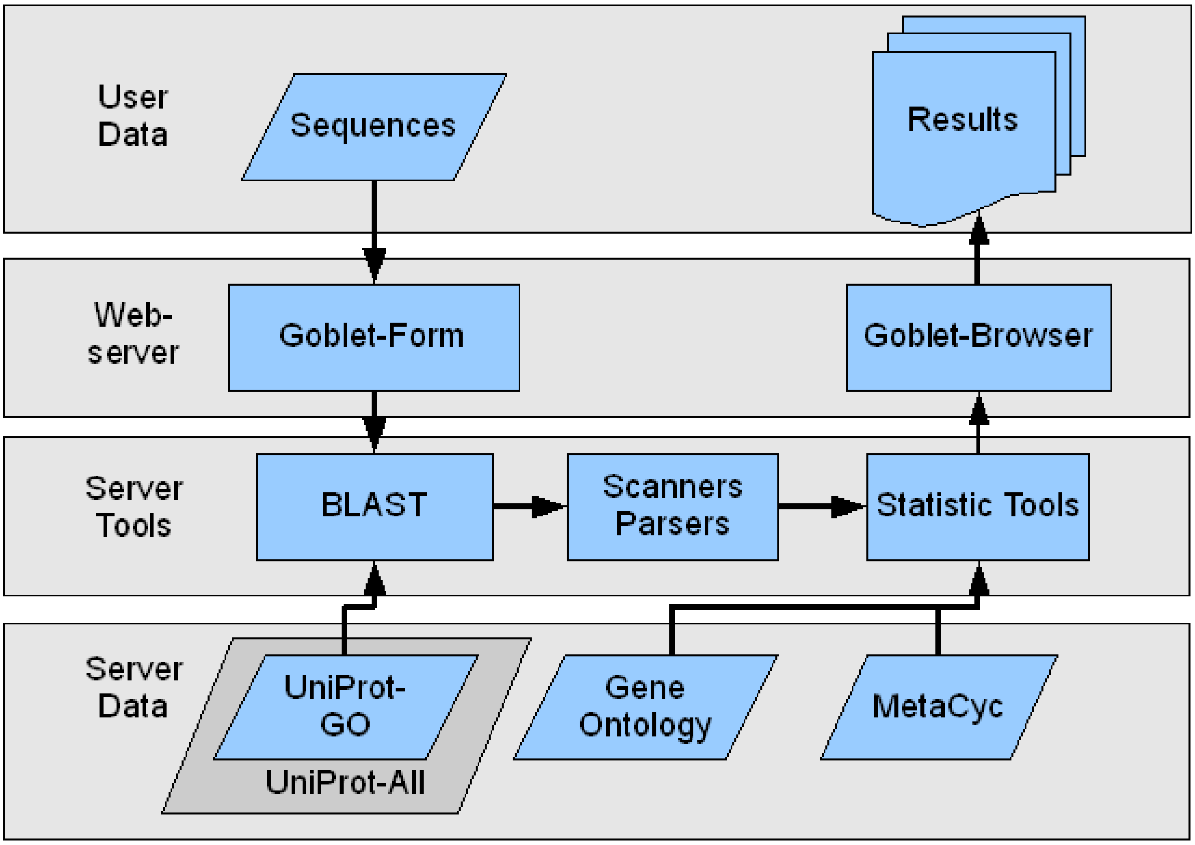


Figure 1: GOblet application schema

Our BLAST databases were generated by downloading the current complete dataset from the UniProt [6] web server and extracting all proteins with GO-annotations. For each UniProt file, two BLAST databases were created, one database containing all proteins that were annotated with GO terms, the other containing only proteins with higher quality annotations that were not inferred from electronical annotations (IEA). By selecting the latter database, users can avoid lower quality annotations and thereby also error propagation of incorrect electronic annotations into their own data sets. The availability of species- and phylum-specific databases allows to restrict searches to the the appropriate search space. The principle structure of the GOblet system is shown in Figure 2.1.

After extraction of GO-annotated proteins, the data are formatted to make them accessible to a local BLAST server. Data for GO, EC-, and MetaCyc-mappings are downloaded from the Gene Ontology website `http://www.geneontology.org` and, to facilitate later steps, converted into a SQLite-database (`http://www.sqlite.org`).

database	all	go(all)	go(-IEA)	MF	BP	CC
sprot-archaea	14017	10273	19	381	156	7
sprot-fungi	20772	12031	9129	1413	1619	534
sprot-human	18886	11318	10119	2021	2636	590
sprot-invertebrates	14730	5833	4474	823	1809	333
sprot-mammals	17767	4993	1310	876	902	251
sprot-plants	24488	10948	1203	566	362	104
sprot-rodents	23660	10398	7711	1740	3072	504
sprot-vertebrates	13310	2827	1271	513	848	182
sprot-viruses	12127	2131	68	75	21	28
trembl-archaea	112996	59134	1	1230	471	110
trembl-fungi	327207	154039	673	1485	704	222
trembl-human	52020	32226	812	1112	812	252
trembl-invertebrates	591970	295444	11181	1818	1949	433
trembl-mammals	65467	49520	354	1061	933	235
trembl-plants	487910	271218	2171	1323	966	256
trembl-rodents	64580	49714	24024	1762	2968	508
trembl-vertebrates	198804	156329	2281	1209	1316	290
trembl-viruses	589940	433856	29	353	259	74

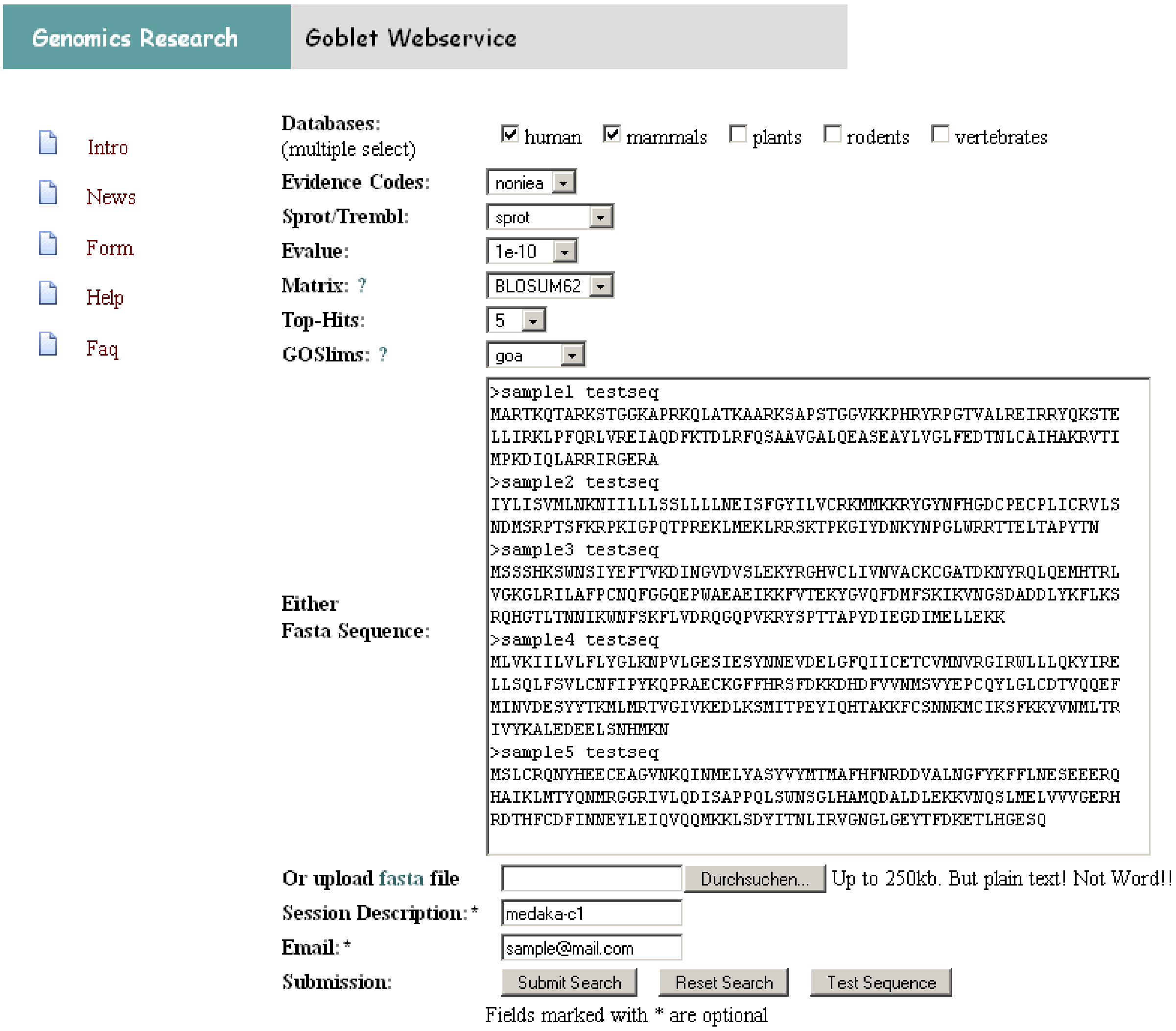
Table 2: GOblet Databases (April 2008). Number of proteins, GO annotated proteins (go), GO annotated proteins without electronically annotated ones (go(-IEA)), and number of GO terms from all annotated proteins in the different GO categories molecular function (MF), cellular component (CC) or biological process (BP).

For the reliable and fast analysis of large data sets, sophisticated database back-ends are generally needed. We are using SQLite for this purpose. SQLite implements a large subset of standard SQL while avoiding the maintenance and configuration overhead of larger systems. The SQLite library is a small C-library which can be easily embedded into other applications. The database itself is just a single file which can be copied, moved and used like any other file. Although many different programming language bindings for SQLite are available, the only scripting language directly supported is Tcl. This fact greatly simplifies the update mechanism for Tcl-applications. By utilizing the Starkit technology (`http://www.equi4.com/tclkit`), several shared libraries for different platforms can be combined into a single executable file. As an example the current GOblet application contains shared SQLite libraries for Windows, Linux, Solaris and MacOSX, which allows to run the application unchanged on these platforms. Libraries for additional platforms can be made available upon request. Tools working on the server are BLAST (which is used to make sequence comparisons), a C-coded BLAST scanner (which converts the BLAST data into a SQLite database), and parsers and statistical tools programmed in the Tcl programming language. Although it would be desirable to utilize other more sophisticated statistic software like the R-programming language [17] here, this would introduce unwanted dependencies and make the implementation of standalone applications more difficult.

GOblet runs as a separate CGI process inside the web server. We are currently using a standard Apache web server (`http://httpd.apache.org/`) on our processing server. For the downloadable version, a Tcl web server (`http://www.tcl.tk/software/tclhttpd`) is embedded into the GOblet application.

The user interface has undergone major changes during the last few years. The Java-Thinlet widget set has now been replaced with standard HTML-code and JavaScript-components. Recent improvements in browser standardization and the implementation of so called widget libraries has greatly facilitated this update, and additional technology like Java is not needed any more inside the browser. All components, except the BLAST tools, are bundled together as a downloadable application which can be run on the users own infrastructure. The GOblet tools and source code are released under the GPL license version 3.

2.2 Annotation pipeline



Databases: (multiple select) ☑ human ☑ mammals □ plants □ rodents □ vertebrates

Evidence Codes: noniea

Sprot/Trembl: sprot

Evalue: 1e-10

Matrix: ? BLOSUM62

Top-Hits: 5

GOSlims: ? goa

Either Fasta Sequence:

```
>sample1 testseq
MARTKQTARKSTGGKAPRKQLATKAARKSAPSTGGVKKPHRYRPGTVALREIRRYQKSTE
LLIRKLPFQRLVREIAQDFKTDLRFQSAAVGALQEASEAYLVGLFEDTNLCAIHAKRVTI
MPKDIQLARRIRGERA
>sample2 testseq
IYLISVMLNKNIILLLSSLLLLNEISFGYILVCRKMMKKRYGYNFHGDCPECPLICRVLS
NDMSRPTSFKRPKIGPQTPREKLMEKLRRSKTPKGIYDNKYNPGLWRRTTELTAPYTN
>sample3 testseq
MSSSHKSWNSIYEFTVKDINGVDVSLEKYRGHVCLIVNVACKCGATDKNYRQLQEMHTRL
VGKGLRILAFPCNQFGGQEPWAEAEIKKFVTEKYGVQFDMFSKIKVNGSDADDLYKFLKS
RQHGTLTNNIKWNFSKFLVDRQGQPVKRYSPTTAPYDIEGDIMELLEKK
>sample4 testseq
MLVKIILVLFLYGLKNPVLGESIESYNNEVDELGFQIICETCVMNVRGIRWLLLQKYIRE
LLSQLFSVLCNFIPYKQPRAECKGFFHRSFDKKDHDFVVNMSVYEPCQYLGLCDTVQQEF
MINVDESYYTKMLMRTVGIVKEDLKSMITPEYIQHTAKKFCSNNKMCIKSFKKYVNMLTR
IVYKALEDEELSNHMKN
>sample5 testseq
MSLCRQNYHEECEAGVNKQINMELYASYVYMTMAFHFNRDDVALNGFYKFFLNESEEERQ
HAIKLMTYQNMRGGRIVLQDISAPPQLSWNSGLHAMQDALDLEKKVNQSLMELVVVGERH
RDTHFCDFINNEYLEIQVQQMKKLSDYITNLIRVGNGLGEYTFDKETLHGESQ
```

Or upload fasta file Up to 250kb. But plain text! Not Word!!

Session Description: * medaka-c1

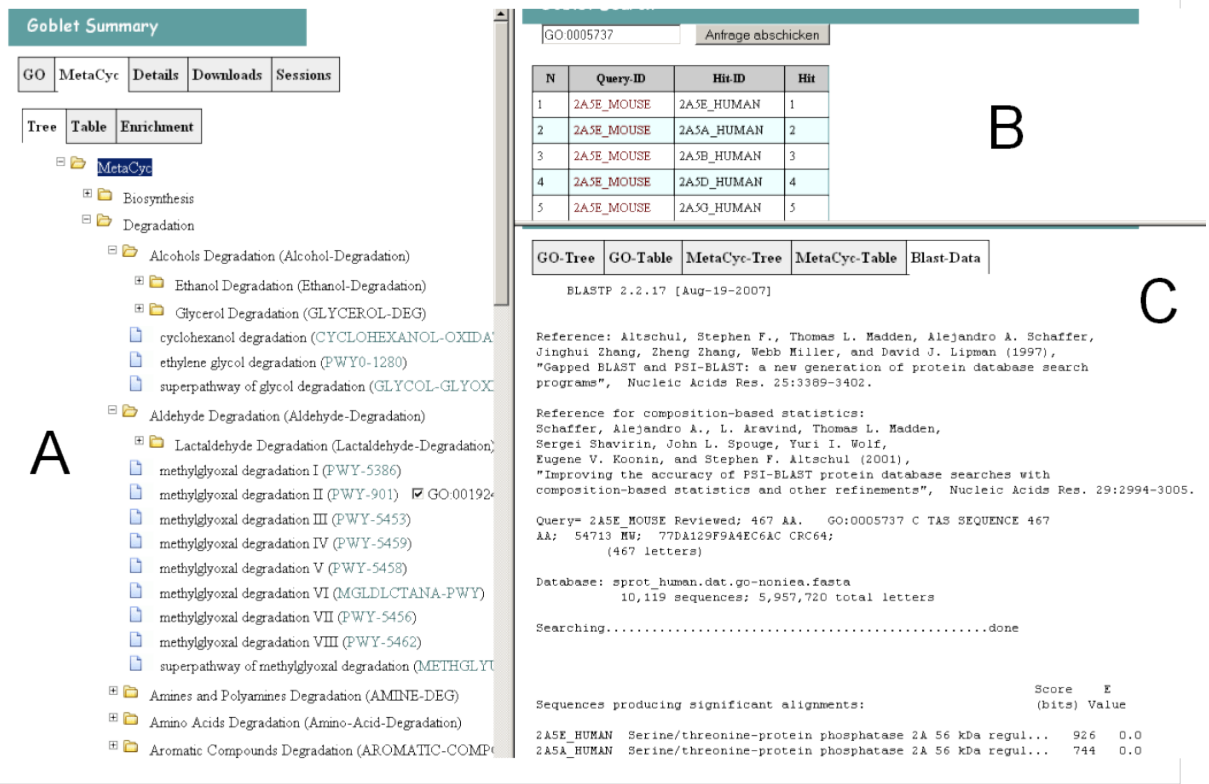
Email: * sample@mail.com

Submission:

Fields marked with * are optional

Figure 2: GOblet sequence submission web form

The advantage of web-enabled applications is that no software packages have to be installed and maintained locally on different platforms. Annotating sequences using the GOblet web server is easy and begins with a user pasting DNA or protein sequence into the web form (Figure 2.2). Alternatively, larger files up to 250 kb, which corresponds to approximately 500 sequences of 500 characters in length, can be uploaded for a single user session. This limit can be adjusted for individual users upon request. The user selects the appropriate data sets and chooses if either all or only non-electronical generated annotations should be used. It is also possible to



MetaCyc
Biosynthesis
Degradation
Alcohols Degradation (Alcohol-Degradation)
Ethanol Degradation (Ethanol-Degradation)
Glycerol Degradation (GLYCEROL-DEG)
cyclohexanol degradation (CYCLOHEXANOL-OXIDA
ethylene glycol degradation (PWY0-1280)
superpathway of glycol degradation (GLYCOL-GLYOX
Aldehyde Degradation (Aldehyde-Degradation)
Lactaldehyde Degradation (Lactaldehyde-Degradation
methylglyoxal degradation I (PWY-5386)
methylglyoxal degradation II (PWY-901) ☑ GO:00192
methylglyoxal degradation III (PWY-5453)
methylglyoxal degradation IV (PWY-5459)
methylglyoxal degradation V (PWY-5458)
methylglyoxal degradation VI (MGLDLCTANA-PWY)
methylglyoxal degradation VII (PWY-5456)
methylglyoxal degradation VIII (PWY-5462)
superpathway of methylglyoxal degradation (METHGLYU
Amines and Polyamines Degradation (AMINE-DEG)
Amino Acids Degradation (Amino-Acid-Degradation)
Aromatic Compounds Degradation (AROMATIC-COMP

GO:0005737

N	Query-ID	Hit-ID	Hit
1	2A5E_MOUSE	2A5E_HUMAN	1
2	2A5E_MOUSE	2A5A_HUMAN	2
3	2A5E_MOUSE	2A5B_HUMAN	3
4	2A5E_MOUSE	2A5D_HUMAN	4
5	2A5E_MOUSE	2A5G_HUMAN	5

```
    BLASTP 2.2.17 [Aug-19-2007]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Reference for composition-based statistics:
Schaffer, Alejandro A., L. Aravind, Thomas L. Madden,
Sergei Shavirin, John L. Spouge, Yuri I. Wolf,
Eugene V. Koonin, and Stephen F. Altschul (2001),
"Improving the accuracy of PSI-BLAST protein database searches with
composition-based statistics and other refinements",  Nucleic Acids Res. 29:2994-3005.

Query= 2A5E_MOUSE Reviewed; 467 AA.   GO:0005737 C TAS SEQUENCE 467
AA;  54713 MW;  77DA129F9A4EC6AC CRC64;
         (467 letters)

Database: sprot_human.dat.go-noniea.fasta
           10,119 sequences; 5,957,720 total letters

Searching..................................................done



                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

2A5E_HUMAN  Serine/threonine-protein phosphatase 2A 56 kDa regul...   926   0.0
2A5A_HUMAN  Serine/threonine-protein phosphatase 2A 56 kDa regul...   744   0.0
```

Figure 3: GOblet result window, summary frame (A), search frame (B), and detailed result frame (C)

extend the data set to be searched by combining several databases. This allows searching, for instance, against the human, rodent, and mammalian data sets simultaneously. Furthermore, the threshold E-value, the number of top BLAST hits to be used for annotation, and the BLOSUM substitution matrix [18] can be chosen by the user. Optionally, a session description and a valid Email address for notification upon job completion can be provided. Immediately after submission, a valid URL is returned. The URL can be bookmarked to retrieve the results. Currently, data are stored for at least ten days on the server.

After job completion, the results are displayed in the browser window within three frames, see the screenshot shown in Figure 2.2. On the left is the summary frame where results about all submitted sequences are displayed (A). In this area the user can explore the Gene Ontology and the MetaCyc tree and select sequences that are annotated with terms of interest for display in the upper right area. The summary area furthermore displays the results of the enrichment analysis for the main GO slims in a tabular format, suitable for pasting the data into a spreadsheet processor. The latter analysis is performed by comparison of the annotated data set against the selected UniProt reference sets using Fisher's Exact Test. All result files can be downloaded as a packed zip-file to enable further local analysis and in order to manually merge different GOblet sessions by the user. The upper right frame is used to display lists of sequence identifiers obtained by either creating queries by the result window or by directly entering search queries in this frame itself (B). Boolean operators can be used to refine queries formerly run against the result dataset. For instance queries like show entries which are involved in ”aging” (GO:0007568) and have ”transcription regulator activity” (GO:0030528) can be shown. The lower right frame is used to display the analysis results for each individual sequence in a tab interface where the annotation results, together with their evidence codes and the underlying BLAST data, can be inspected (C).

3 Example Use

In order to test the reliability of annotations made by GOblet, we selected a reference data set containing 1045 mouse proteins from the sprot rodents database released by UniProt. The aim was to test for how many mouse proteins we could get their original TAS annotation by GOblet (re-assignment). We selected proteins that have GO annotations with the highest ranked evidence code "traceable author statement" (TAS). There were a total of 1804 annotations with this evidence code for mouse proteins available. We annotated this data set with the GOblet web server by choosing the human sprot-database with only higher ranking evidence codes, i.e., where IEA codes are excluded. A cutoff E-value was 1e-10, and the substitution matrix BLOSUM80 was chosen. We used for the annotation the five best hits from the BLAST result. Using this settings, we could annotate 991 mouse proteins, 636 of which matched exactly at least one of its TAS verified gene ontology term (61%). From the annotations we could re-assign 972 from 1804 annotations for the right protein (54%). In a second experiment we chose an E-value of 1e-50 and took only the three top BLAST hits for the annotation. Here we could re-assign 563 proteins and 972 annotations. By increasing the evidence stringency in this manner, the number of total mappings between mouse proteins and gene ontology terms was reduced from 11006 to 7108. It should be noted that the number of total mappings was reduced by about 35 %; the number of re-assigned proteins just dropped by about 12 %.

A larger amount of the GO terms could not be re-assigned. The reason for this might be the fact that the annotation of human proteins with GO terms currently is not complete. Some examples demonstrate that the annotation quality is only as good as the underlying database which is used for the annotation. The protein PTGES_HUMAN (Prostaglandin E synthase) has no annotated molecular function, although there exists a GO term prostaglandin-E synthase activity (GO:00050220). Therefore, the molecular function for the mouse protein PTGES_MOUSE could not be assigned. However, as could be seen from the UniProt database entry, it is annotated with the EC number 5.3.99.3., which stands (according to the external2go mappings provided from the Gene Ontology website) again for "prostaglandin-E synthase activity". Therefore, if the EC numbers were used together with the GO annotations, this protein could be annotated correctly.

A second example where the original GO term could be not assigned is the mouse protein Angiopoietin-2 precursor (ANGP2_MOUSE), which matches its ortholog ANGP2_HUMAN. However, as the mouse protein is annotated more specific with "vascular endothelial growth factor receptor binding" (GO:0005172), its human counterpart is less specific annotated with "receptor binding" (GO:0005102). Another example is the Synaptonemal complex protein 1 (SYCP1_MOUSE) which just has one annotation in the cellular component ontology with quality TAS: "lateral element" (GO:0000800). Unfortunately, its human counterpart SYCP1_HUMAN is currently not annotated in the cellular component ontology with non-IEA codes.

Further examples where the GOblet annotation pipeline has been useful for us are experiments where we would like to get an overview about possible candidates of genes carrying interesting functions selected for downstream processing. In a related study by Hong and Coworkers [19] GOblet has been successfully used for the identification and integrative analysis of novel genes specifically expressed and developmentally regulated in murine spermatogenic cells. For us the enrichment analysis proved to be useful also in a recent analysis in which we could correlate molecular and morphological data. Gene Ontology annotations using GOblet has been used as well for creating a web interface providing function oriented exploration of genomes (http://www.molgen.mpg.de/~ag_seaurchin).

4 Discussion and Conclusions

We are introducing a major update of the GOblet web service, and many of the new and improved features were implemented based on users' feedback. We are now offering both a web service, i.e., no software needs to be installed on the client machine, as well as a stand-alone application, which can be used to annotate high-throughput data and/or confidential data. Furthermore, the addition of pathway annotations allows a more integrative sequence analysis based on MetaCyc terms, thus avoiding some of the pitfalls associated by using the GO terminology alone. Finally, statistical analyses of annotations allow a more powerful and rigorous examination of the results. These analyses include automatic detection of significantly enriched ontology and pathways terms as well as K-means [20] clustering of sequences according to their GO annotation.

In addition to the annotation for individual sequences, a summary about the results is usually required. This includes grouping related sequences together and performing statistics about the main features of the data set, such as reporting the most abundant terms, and the possible enrichment of such terms against the expected frequencies from reference data sets. Furthermore, it might be valuable to structure, or cluster, the data together into groups that share the same annotation terms. In contrast to our approach of simply using a binary matrix of a GO terms as columns and sequences as rows, it might be of interest to measure the semantic similarity between GO terms. For example the method developed by Lord and co-workers [21] could be used. Here the similarity is considered as a linkage of terms inside the GO tree, weighted by a probability distribution that balances the non-uniform "richness" of different parts of the direct acyclic graph. In future versions, we might consider implementing approaches like those of Pozo and co-workers [22], in which functional similarities between GO terms are quantified by utilizing the co-occurrence of GO-terms in the same set of InterPro identifiers. Such a strategy would avoid the sometimes artificial distinction between different branches in the GO hierarchy made by human curators.

From our example cases it can be concluded that the completeness of the GO mappings in the UniProt database is the main point of successful annotation. Possible solutions to missing annotations are for instance the transfer of GO annotated EC identifiers into the database or the usage of Swiss-Prot keywords, which are part of the protein annotation, to get at least a minimal although unspecific annotation. Of course many more mappings could be utilized but it should be noted that the recently observed explosion of IEA go annotations on datasets might be problematic.

Future developments will include improved data management facilities on the web server, where users can store, combine, and analyse their data directly on the server. We also are in the process of implementing a SOAP-web service that is suitable for inclusion into tools connecting web service tools like Taverna [23].

With the availability of a downloadable version of the GOblet system users can now setup their own high-throughput pipelines. The advantages include that maintainers can follow their own update cycles, integrate their own databases, and are independent of the actual performance or availability of the main GOblet web server. By getting access to the source code, experienced programmers can actually improve and extend the system.

Although the integration of pathway information and Gene Ontology terms provides a significant improvement over previous GOblet versions, it cannot yet take into consideration orthology relationships of sequences for the purpose of functional annotation. To this end, incorporating phylogenomics approaches [24, 25] might prove to be a powerful addition to future GOblet versions. Similarly, it needs to be carefully evaluated whether the inclusion of information about sequence patterns [26, 27] or the integration of other existing strategies for the annotation of protein sequences [28] will prove to complement and improve our current approach.

Acknowledgements

The authors acknowledge Patrick May and Björn Usadel for fruitful discussions.

References

[1] J. Quackenbush. Extracting biology from high-dimensional biological data. *Journal of Experimental Biology*, 210:1507–1517, 5 2007.

[2] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[3] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S. A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25:1251–1255, 2007.

[4] Gene Ontology Consortium. The Gene Ontology project in 2008. *Nucleic Acids Research*, 36:D440–444, 1 2008.

[5] N. Daraselia, A. Yuryev, S. Egorov, I. Mazo, and I. Ispolatov. Automatic extraction of gene ontology annotation and its correlation with clusters in protein networks. *BMC Bioinformatics*, 8:243, 2007.

[6] H. Wu, Z. Su, F. Mao, V. Olman, and Y. Xu. Prediction of functional modules based on comparative genome analysis and gene ontology application. *Nucleic Acids Research*, 33(9):2822–2837, 2005.

[7] S. Draghici, P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, 2 2003.

[8] A. Kanapin, S. Batalov, M. J. Davis, J. Gough, S. Grimmond, H. Kawaji, M. Magrane, H. Matsuda, C. Schönbach, R. D. Teasdale, and Z. Yuan. Mouse proteome analysis. *Genome Research*, 13:1335–1344, 6 2003.

[9] W. Kusnierczyk. Taxonomy-based partitioning of the gene ontology. *Journal of Biomedical Informatics*, 41(2):282–292, 4 2008.

[10] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36:D480–D484, 2008.

[11] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(Database issue):D428–D432, 2005.

[12] R. Caspi, H. Foerster, C. A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Y. Rhee, A. G. Shearer, C. Tissier, T. C. Walk, P. Zhang, and P. D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 36:D623–631, 1 2008.

[13] X. Mao, T. Cai, J. G. Olyarchuk, and L. Wei. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics*, 21:3787–3793, 10 2005.

[14] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability & Statistics. John Wiley and Sons Inc., 15 edition, 2002.

[15] D. Groth, H. Lehrach, and S. Hennig. GOblet: a platform for Gene Ontology annotation of anonymous sequence data. *Nucleic Acids Research*, 32:W313–317, 7 2004.

[16] S. Hennig, D. Groth, and H. Lehrach. Automated Gene Ontology annotation for anonymous sequence data. *Nucleic Acids Research*, 31:3712–3715, 7 2003.

[17] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008.

[18] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 11 1992.

[19] S. Hong, I. Choi, J. M. Woo, J. Oh, T. Kim, E. Choi, T. W. Kim, Y. K. Jung, D. H. Kim, C. H. Sun, G. S. Yi, E. M. Eddy, and C. Cho. Identification and integrative analysis of 28 novel genes specifically expressed and developmentally regulated in murine spermatogenic cells. *Journal of Biological Chemistry*, 280(9):7685–7693, 3 2005.

[20] J. A. Hartigan and M. A. Wong. A k-Means clustering algorithm. *Applied Statistics*, 28(1):100–108, 1979.

[21] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19:1275–1283, 7 2003.

[22] A. del Pozo, F. Pazos, and A. Valencia. Defining functional distances over gene ontology. *BMC Bioinformatics*, 9:50, 2008.

[23] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. R. Pocock, P. Li, and T. Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, 34:W729–732, 7 2006.

[24] J. A. Eisen. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research*, 8:163–167, 3 1998.

[25] B. E. Engelhardt, M. I. Jordan, K. E. Muratore, and S. E. Brenner. Protein molecular function prediction by Bayesian phylogenomics. *PLoS Computational Biology*, 1(5):e45, 10 2005.

[26] E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. InterProScan: protein domains identifier. *Nucleic Acids Research*, 33:W116–120, 7 2005.

[27] N. J. Mulder, R. Apweiler, Attwood, et al. InterPro, progress and status in 2005. *Nucleic Acids Research*, 33:D201–205, 1 2005.

[28] I. Friedberg. Automated protein function prediction–the genomic challenge. *Briefings in Bioinformatics*, 7(3):225–242, 9 2006.